DOCUMENT RESUME

ED 307 290 TM 013 279

AUTHOR Olson, George H.
TITLE On the Validity of Performance Grades: The Relationship between Teacher-Assigned Grades and Standard Measures of Subject Matter Acquisition.
PUB DATE Mar 89
NOTE 19p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, March 28-30, 1989).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; Achievement Rating; Comparative Analysis; Correlation; Evaluation Methods; Grades (Scholastic); *Grading; High School Students; Performance; Secondary Education; *Secondary School Teachers; *Standardized Tests; Student Evaluation; Teacher Made Tests; *Validity
IDENTIFIERS Final Examinations; Subject Content Knowledge; Survey Tests Essential Elements Learner Standards; Texas

ABSTRACT

The validity of grades given by secondary school teachers was studied. Two teacher-originated measures of performance and one independent standardized measure of achievement were examined. The teacher-originated grades were final examination grades (FINALs) and end of the semester marks (MARKs) for over 40 courses at all secondary levels. The standardized tests were from the "Survey Tests of Essential Elements/Learner Standards" (STEELS) covering the instructional content of the more than 40 courses studied, spanning the areas of language arts, mathematics, social studies, and science. Over the 44 tests examined, correlations among the three measures--MARKS, FINALS, and STEELS--often varied markedly among each other both within and among tests. No strongly consistent patterns were observed; only weak trends and tendencies were found. Within-teacher and between-school validity coefficients tended to be greatest in freshman level courses with large enrollments, which were heterogeneous in competency. There were no clear differences in marking and grading across disciplines. The generally low validity coefficients for teacher final examinations and end of semester grades imply that these measures reflect many characteristics of students. An example of the marking and grading systems typical of those used by teachers in the Dallas school system (Texas) illustrates the difficulties of marking systems, particularly when the student receives a zero for incomplete work. Ten tables provide data on student performance. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED307290

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
GEORGE H. OLSON
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

On the Validity of Performance Grades: The Relationship Between Teacher-Assigned Grades and Standard Measures of Subject Matter Acquisition

George H. Olson
Dallas Independent School District

March 1989

The marks students receive on their report cards are among the most influential determinents of their current and future behavior they will ever encounter. Grades play a crucial role in children's personal, academic, and future professional lives. Most immediately, the grades young people receive influence the way in which they are perceived by their peers, their teachers, and their parents. Even more importantly, their grades can affect their own perceptions of self worth. Poor grades can lead to feelings of despondency, worthlessness, and alienation. Good grades, on the other hand, can often inspire higher levels of confidence, motivation, and industry.

In more concrete terms, grades play a role in such mundane matters as how much allowance a child receives, whether or not he or she has access to the family car (or to his or her own car), and the lateness of his or her curfew. In more important matters, grades determine class rank and honors, and can influence athletic elligibility, access to employment, the level of placement once employed, and the likelihood of college acceptance.

Yet, in spite of their importance, marking and grading systems are characterized by inconsistency and inaccuracy. While most parents, college registrars, and potential employers view grades as indicative of performance and attainment of competence, teachers use grades for a variety of academically related and unrelated purposes including punishment for discipline and control, and reward for industry and effort, general appearance and comportment, and class status.

There is no shortage of literature and research on marking and grading. The vast majority of this work, however, has been carried out in the college and university environment. A comprehensive review of this corpus can be found in the book, Making Sense of College Grades (Milton, Pollio, & Eison, 1986). Another excellent source, well worth reading by anyone interested in the problems associated with college-level grading and marking, is the Psychological Bulletin monograph by Duke (1983). The research reviewed in these sources and elsewhere has yielded a rich array of findings and conclusions. Three that have been particularly pervasive are the

Paper presented at Session B2 of the 1989 annual meeting of the National Council on Measurement in Education in San Francisco, March, 1989.

BEST COPY AVAILABLE

TM013279

following: grades and marks reflect considerably more than the achievement of competence (see, for example, Airasian, 1984; Leiter & Brown, 1985; Pedulla, Airasian, & Madaus, 1980; Salmon-Cox, 1981; and the several references and examples given in Hills, 1981); differential standards for awarding grades and marks exists across disciplines, fields of study within disciplines, and instructors (Duke, 1983; Goldman, Schmidt, Hewitt, and Fisher, 1974; Prather, Smith, & Kodras, 1976); and marking and grading are a relative enterprise in that instructors' distributions of marks tend to be invariant over time and institutions regardless of any real differences or changes in achievement.

Much less research on marking and grading has been conducted at the elementary and secondary school level. Typically, what relevant research does exist has been concerned with teachers' use of tests -- particularly standardized tests (Gullickson, 1984, 1986; Gullickson & Hopkins, 1987; Stiggins, Conklin, & Bridgeford, 1986; and Lazar-Morris, Polin, & Barry, 1980). This research has found teachers ill-prepared to use or develop tests and assessments adequately.

While the research on teachers' practices of marking and grading is sparse, professional commentary has been plentiful (see, for example, the elementary textbooks on measurement and testing by Ebel, 1979; Mehrens & Lehmann, 1984; and Hills, 1981). These authors are genuinely concerned with the poor preparation teachers generally receive prior to entering the teaching profession. Currently there are few if any requirements that potential teachers be trained in testing in order to be certified (Stiggins, Conklin, & Bridgeford, 1986).

The focus of this paper is on the validity of secondary school teachers' assessments; in particular, the study reported here examines the validity of teachers' marks. Any study of validity necessarily begs a question of reference: with respect to <u>what</u> should marks be valid?

Specialists in educational measurement typically advance the notion that the marks students are given by their teachers should reflect, in so far as possible, relatively pure measures of achievement. Thorndike (1969), for example, takes the position that marks ought to represent unequivocal estimates of competence. Taking a similar posture is Hills (1981) who gives the purpose of marking and grading as one of communicating level of achievement free of distortion from other characteristics of students. This is the notion of validity examined here. Thus, the paramount question addressed in this study is: can teachers' marks be interpreted as valid reflections of their students' achievement of competence?

PROCEDURES AND ANALYSIS

<u>Measures of Performance and Achievement</u>

Two teacher-originated measures of performance and one independent, standardized measure of achievement were examined. The teacher-originated measures, subsequently referred to as "performance grades," were the end of the second semester (spring 1987) final exam grades and the end of

semester marks. Both of these measures are, by school district policy, on a functional numerical scale from 50 to 100. The standardized measure of achievement was a battery of locally-constructed curriculum-referenced tests.

Performance Grades. Final exams (FINALs) typically consisted of tests constructed by the teachers themselves. These included anything from multiple choice examinations to essay questions. No information was available on the reliability of these measures. A district policy manual (DISD, 1987) dictated what must be taken into account in determining end of semester marks (MARKs). These were to be composed of five weighted components: each of three six-week cycle grades carried a weight of 25 percent, the FINAL carried a weight of 15 percent, and scores on the locally constructed tests carried a weight of 10 percent.

The policy manual further stipulated that in computing six-week (numerical) cycle grades teachers were to take into consideration homework, class participation, test performance, special projects, and "other assignments as appropriate." Teachers or department heads, at the school's discretion, were free to determine how much weight each of these components would carry.

Standardize Achievement Measures. The standardized measures of achievement were scores from a locally constructed battery of tests aimed at measuring students' progress toward mastery of district-wide instructional objectives. The secondary school Survey Tests of Essential Elements/Learner Standards (STEELS) is composed of more than 40 tests covering the instructional content of over 40 corresponding courses spanning four disciplines: language arts, mathematics, social studies, and science. Development of the STEELS, which is on-going, has taken several years and has followed principles of sound test construction (Grobe, et. al., 1987). Items have been selected that discriminate well between higher and lower achieving students, that cover a range of difficulty wide enough to allow accurate measurement over a broad range of ability, and that are closely matched to the content of district-used textbooks and curriculum materials. This effort has resulted in a set of curriculum-relevant tests with excellent psychometric properties.

Along with the development of the STEELS was a concomitant effort aimed at the development of a suitable interpretive scale on which to express STEELS results. The scale had to have two properties: a cut score representing a minimum passing score; and a functional range similar to that of the other measurement scales used by teachers (i.e., it had to have a functional range of 50 to 100). While the usually recommended procedures for constructing a scale having these properties involve convening a panel of content or curriculum experts to "set" a cut score, we, for various reasons (see Olson, 1989) could not use this approach. Instead, a simple expediency was used in setting the cut score: that raw score on a particular STEELS test at or above which the same percentage of examinees scored as passed the corresponding course the previous year was equated to a scale score of 70. A perfect score on the test was equated to a scale score of 99.9. Using these two points as anchors, and procedures suggested by Angoff (1984), a normalized, percentile derived linear scale was constructed. The scale provided both a normative and a functional interpretation: normative in that an individual's scale score showed where the

student stood relative to other students throughout the district; functional in that a meaningful passing score was established and that differences in scale scores represented interval differences in achievement.

Assessing the Validity of Performance Grades

In addition to providing the intended information relative to the district's schools' and students' progress toward meeting the district's instructional objectives, the STEELS provided a heretofore unavailable, independent basis for examining the validity of the performance grades awarded by teachers. While it can be expected that methods, procedures, and even standards for assigning grades will differ somewhat from teacher to teacher and on average from school to school, we should not expect large differences, for then a grade of 82, say, in a particular course in one school would bear little meaning relative to a similar grade in the same course at another school. Certainly, the district would like to be able to state that a student in School A who received a final grade of 76 in Algebra I had achieved roughly the same amount as a student in School B who received the same grade in Algebra I. In other words, the district would like to be able to say that its teachers' performance grades are valid indicators of achievement.

In this analysis validity of performance grades is evaluated by correlations between FINALs and STEELS and between MARKs and STEELS. In this context, the STEELS are assumed to be a relatively pure measure of students' achievement of the district's curriculum objectives. The validity of performance grades is based upon the degree to which these also measure achievement. Validity, as considered here, however, can be examined from more than one perspective or frame of reference (Thorndike, 1969). On the one hand, we can view the validity of the performance measures from a districtwide frame of reference in terms of how they are used across schools. This requires that the same, absolute standards for marking and grading apply everywhere. On the other hand, validity can be viewed in a more provincial sense as applying only within schools, or even among teachers' own students. This requires only that schools, or teachers within schools, adhere to a given set of standards when awarding marks for achievement. In the analyses reported here, validity was evaluated from three perspectives.

Between Schools. If validity is to be vieved in absolute terms, then teachers' performance grades ought to be awarded on the basis of a common set of achievement standards throughout all schools in the district. When schools differ in mean achievement, the mean performance grades awarded should differ correspondingly. In other words, between school validity, computed as correlations between school mean performance grades (FINALs and MARKS) and mean achievement (STEELS), is required in order for performance grades to have consistent meaning from one school to the next.

Between Teachers Within Schools. Teachers within schools might view achievement in terms of locally-relevant standards. Thus, students in one school's highest achieving classes might receive high performance grades even though they might achieve at a lover level than the students in the highest achieving classes who receive the same performance grades at another school. From this frame of reference high validity coefficients comput-

ed by correlating mean performance grades and achievement over teachers within schools would be expected.

Within Teachers and Schools. From yet another frame of reference validity would be indicated by high correlations between the performance grades individual students receive and their levels of achievement. If teachers were found to be consistent in their application of achievement standards in awarding performance grades to their own students then we could conclude that at least within teachers their marking and grading practices were valid.

RESULTS

A complete accounting of the relationships among the performance measures of all the tests in the STEELS battery and their corresponding courses is well beyond the scope of this paper. That detail is available elsewhere (Meier, 1988; Olson, 1988); here we will examine in detail only a few representative cases. However, before looking at some specific examples, it may be worthwhile to point out some general findings.

Overall Results

Over the 44 tests examined, the correlations among the three measures (MARKs, FINALs, and STEELS) often varied markedly among each other both within tests and among tests. Generally, however, the validity coefficients for the performance grades were considerably lower than the correlations between teacher final exam and end of semester marks. These, as might be expected, were almost uniformly large. Furthermore, although there were important exceptions, the validity coefficients tended to be strongest when computed within teachers then when computed among teachers within schools or among schools. On the other hand, the coefficients computed among schools tended to be higher than those computed among teachers within schools; although some important exceptions were observed: in a few instances schools having STEELS means ranking among the bottom of the distribution were observed to have awarded some of the highest final exam and end of semester grades, resulting in negative validity coefficients.

Generally, an analysis of the data from all the tests and courses lead to the conclusion that teachers tend to award performance grades partially on the basis of relative achievement within their own classes of students. It was just as obvious, from the size of the correlations, that performance grades were based on factors other than achievement. The fact that one class of students may be of measurably higher caliber than another class within the same school appeared to be irrelevant. Furthermore, although many exceptions were noted, there appeared to be a weak trend toward awarding higher performance grades to students in schools having, on average, measurably higher achieving students.

Results From Specific Courses

At the end of this report are several tables summarizing the results of analyses of eight of the courses investigated. These represent a repre-

sentative sampling in that tests from each of the four disciplines are included. Additionally the eight cases represent long tests, short tests, courses in which only a few students were enrolled, courses in which large numbers of students were enrolled, introductory courses, and advanced courses.

The analyses of each course are summarized in two tables. The first provides numbers of items, KR_{20} reliability coefficients, numbers of cases, means, covariances, and correlations among the three measures of performance (final exam grades, end of semester grades, and STEELS scale scores) for each level of analyses (among schools, among teachers, and among students). The second table displays the results of a fully nested analysis of variance for each of the performance measures. These are included to show where reliable differences among kids, teachers, and schools actually existed. Even these tables, which represent only a small portion of the full analysis, contain too much detail to be gone into in depth here. The extensive detail for these eight courses has been included for the benefit of those readers who desire the opportunity to draw their own conclusions.

Between Schools Validity. Lower level (freshman) courses with large enrollments (English I, Algebra I, Biology I, and U.S. History) tended to show the greatest differentiation among school means on all three performance measures (FINALs, MARKs, and STEELS). For these courses, the between school validity coefficients are significant and fairly strong, especially in light of the fact that variance among school means, in general, tends to be severely attenuated. Teachers of courses in schools having higher achieving students on average awarded the highest performance grades. This was particularly evident in the two courses (English I and Algebra I) that are offered as pre-honors courses for advanced students in junior high schools. Some summary data from these two courses for a selected set of schools are shown in Tables 1 and 2. In each table, the schools were selected such that the lowest and highest mean for each of the three measures would be represented. In Table 1 the last three schools are junior high schools. In Table 2, the last two schools are junior high schools. It is obvious from the two tables that most of the between school variance on all three measures was due to the distinction between junior and senior high schools. Advanced, highly achieving students take these two courses early and are rewarded with higher performance grades.

Table 2, incidently, points up one of the factors that weakens the between school validity of performance grades. School 8, which barely averaged a passing STEELS score exhibited the highest end of semester marks by a wide margin. On the other hand, School 24 had a mean STEELS score that was definitely higher than the other three high schools shown yet awarded the lowest final exam scores and end of semester marks.

In the more advanced (senior level) courses (English IV, Pre-Calculus, World Geography, and Introduction to Physical Science) the relationship between performance grades and achievement was not nearly as strong. In fact in several instances, most notably in World Geography and Introduction to Physical Science, the correlations between performance grades and achievement could not be significantly distinguished from zero.

Table 1

Means and Within School Correlations for
English I in Six Schools

School	Mean Performance Measure FINAL	STEELS	MARK	w/ School Correlations FINAL, STEELS	FINAL, MARK	STEELS, MARK
3	76.7	71.6	73.4	.44	.71	.54
18	64.1	72.8	67.9	.29	.81	.47
23	65.1	76.4	71.2	.49	.68	.61
53	90.6	85.9	87.9	.55	.60	.62
58	85.6	88.1	88.1	.06	.52	.14
60	97.3	77.2	88.7	.98	.47	.49

Note: The schools displayed represent the highest and lowest means on each of the three measures of performance.

Table 2

Means and Within School Correlations for
Algebra I in Five Schools

School	Mean Performance Measure FINAL	STEELS	MARK	w/ School Correlations FINAL, STEELS	FINAL, MARK	STEELS, MARK
3	76.9	67.7	73.8	.37	.75	.51
8	73.2	70.1	96.6	.23	.62	.50
24	67.7	76.6	70.7	.47	.84	.66
53	85.0	94.3	89.3	.31	.70	.55
60	91.9	84.4	81.3	.08	.68	.55

Note: The schools displayed represent the highest and lowest means on each of the three measures of performance.

Between Teachers Within Schools Validity. Among teachers, validity was virtually nonexistent. Even though there were significant differences among means of teachers' students on all three performance measures, these differences were not correlated. In Pre-Calculus and World Geography the correlations between performance grades and STEELS were not significant. In Algebra I, U.S. History, Biology I, and Introduction to Physical Science, less than ten percent of the variance was common among either of the performance grades and STEELS. In the English courses, the validity coefficients were a little stronger, but not by much. Apparently teachers make little distinction among the relative achievement of each other's students when awarding performance grades.

Within Teachers Validity. The strongest within teacher validity coefficients were observed for FINALs and MARKs in English I, Algebra I, and U.S. History; and for FINALs in Biology I and Pre-Calculus. While English I and Algebra I share the characteristic of being offered as honors courses in junior high school they shared little in common with U.S History (a freshman survey course) and nothing at all with Biology and Pre-Calculus (advanced level courses). An explanation of why these courses should have higher within teacher validity than the others cannot be given. The explanation does not appear to be related to the psychometric characteristics of the STEELS tests, the level of the course, the numbers of students enrolled, or the heterogeneity or homogeneity or the students.

Comparison of Validity Coefficients. Except for U.S. History and World Geography, there was a strong tendency for the correlation between final exam and STEELS score to be stronger than that between end of semester grade and STEELS score. This may be due to the final exams being more similar to the STEELS in that both were tests of achievement. The fact that these correlations were not higher implies problems with the psychometric quality of the teachers' tests. The lower correlations with end of semester grades is more likely than not due to the independent variance associated with several of the components that go into the calculation of the end of semester grade.

SUMMARY AND CONCLUSION

The results of the analysis of validity of performance grades in over forty courses were somewhat of a mixed bag. No strongly consistent patterns were discovered, only weak trends and tendancies. Thus, within-teacher and between-school validity coefficients tended to be greatest in freshman level courses with large, heterogeneous (in competency) enrollments; although, numerous exceptions existed.

Several instances of differential grading standards among teachers and schools were noted. Again, however, no consistent patterns could be descerned, as have been found in studies in higher education. There were, for instance, no clear differentials in marking and grading across disciplines.

On the other hand, the generally low validity coefficients for teacher final exams and end of semester grades imply that these measures reflect many characteristics, of students and otherwise, that may be independent of achievement. There are many candidates for what these might be. For

instance, as Notingham (1988) recently pointed out, it is politically impossible to flunk too many students. Thus, teachers adopt such practices as allowing students to retake tests on which they performed poorly, providing assignments for extra credit, and altering the difficulty level of tests, often unwittingly, from one testing occasion to the next. Additionally, while good-intentioned, teachers are as apt to grade on the basis of socially desirable characteristics, and on motivation and effort, as they are on achievement and competence (Airasian, 1984; Barns, 1985).

The low validity of teacher final examinations probably reflects the insufficient training in test construction teachers receive prior to entering the profession (Carter, 1984; Farr & Griffin, 1973; Fleming & Chambers, 1983; Gullickson & Ellwein 1985). The low validity of semester grades is indicative of a larger problem: teachers are ill-prepared for the complex processes involved in evaluating and grading students. Even when the components of a semester grade are prescribed, as they are in the Dallas school system, teachers still find enough leeway to vary grades independently of student achievement.

There is a considerable need for prospective teachers to be better trained in procedures for student evaluation and assessment. Several years ago, Farr and Griffin (1973) reported that teachers had not been taught what they needed to know about measurement and evaluation of students before entering the classroom.

Furthermore, even though they had limited knowledge of measurement principles, and considered the practice of grading and marking to be one of the most difficult aspects of teaching (Barns, 1985), Farr and Griffin found little inclination among teachers to pursue this knowledge after entering the profession. Farr and Griffin suggested that the reason for teachers' lack of interest in measurement was a result of the existential way in which measurement, as a discipline, is taught to prospective teachers. They contended that teachers are not interested (and probably have no need to be interested) in the quantitative aspects of measurement, as it is typically taught. Rather, teachers need to come to view measurement as a purposeful activity inherent in the whole process of teaching.

The problems with the differences between what is taught (of measurement concepts and principles) and what is needed were delineated quite eloquently over 45 years ago when Scates (1943) outlined the incompatability between what the measurement specialist demanded of (scientific) measurement and what the classroom teacher required for the on-going, day-to-day, <u>in-situ</u> evaluation of students. (Scates' paper, which has been extensively quoted in Stiggins, Conklin, and Bridgeford (1986) is well worth reading in its entirety.) Several years later, Thorndike (1969), after completing his review on marking and marking systems, concluded that the literature on marking and grading over the previous 50 years haD "missed the mark" in that it failed to appreciate the "very real limits in time, precision of judgement, and skill in assessment within which the typical teacher operates." More recently, Stiggins, Conklin, and Bridgeford (1986) have expressed this same sentiment.

TABLE 3

Means, Correlations, and Covariances for Test No. 10
English I

(Number of Items = 60; KR_{20} = .88)

	Between Schools (N = 50)			Teachers w Schools (N = 404)			Pooled w. Schools (N = 6366)		
	FINAL	STEELS	MARK	FINAL	STEELS	MARK	FINAL	STEELS	MARK
	78.08	78.71	78.63	72.98	75.80	73.87	74.96	76.73	75.04
CORRELATION \ COVARIANCE MATRIX[a,b]									
FINAL	30.48	19.15	30.20	94.82	26.09	71.13	74.96	76.73	75.04
STEELS	.72	23.06	22.24	.46	34.12	22.31	.58	49.58	42.64
MARK	.93	.79	34.46	.79	.41	85.79	.77	.54	127.57

[a]Variances are given on the diagonals, covariances above the diagonal, and correlations below the diagonal.

[b]All correlation coefficients are significant at $p < .01$.

TABLE 4

Nested Analysis of Variance of Final Test Grade, STEELS Scores, and Course Marks for Test 10: English I

(Kids w/ Teachers w/ Schools)

		FINAL		STEELS		MARK	
Source	d.f	M. Sqr.	F	M. Sqr.	F	M. Sqr.	F.
Schools (S)	49	2409.76	3.3	1882.96	5.4	2722.85	4.0
Tchrs. (T) w S	354	729.30	3.9	347.64	4.5	681.21	3.7
Kids w T w S	5962	185.68		76.52		183.44	

Note: All F-Ratios are significant at $p < .001$.

TABLE 5

Means, Correlations, and Covariances for Test No. 13
English IV

(Number of Itmes = 85; KR_{20} = .88)

	Between Schools (N = 26)			Teachers w Schools (N = 147)			Pooled w. Schools (N = 2427)		
	FINAL	STEELS	MARK	FINAL	STEELS	MARK	FINAL	STEELS	MARK
	78.02	77.25	79.39	77.56	76.86	79.78	78.51	77.61	79.87
	CORRELATION \ COVARIANCE MATRIX[a,b]								
FINAL	13.58	4.15	7.97	74.21	16.39	44.15	86.24	27.53	56.79
STEELS	.38	8.76	2.20	.39	24.18	12.59	.49	36.86	21.10
MARK	.73	.25	8.79	.76	.34	55.12	.70	.34	76.38

[a]Variances are given on the diagonals, covariances above the diagonal, and correlations below the diagonal.

[b]Underlined correlation coefficents are not significant; all others are significant at at least $p < .05$.

TABLE 6

Nested Analysis of Variance of Final Test Grade, STEELS Scores, and Course Marks for English IV

(Kids w/ Teachers w/ Schools)

		FINAL		STEELS		MARK	
Source	d.f	M. Sqr.	F	M. Sqr.	F	M. Sqr.	F.
Schools (S)	25	965.75	1.9	722.04	2.5	585.21	1.7
Tchrs. (T) w S	121	500.14	4.0	289.27	2.9	354.25	3.4
Kids w T w S	2280	123.92		99.40		105.10	

Note: The between school F-Ratios for FINAL and MARK are significant at $p < .01$; the other F-Ratios are significant at $p < .001$.

TABLE 7

Means, Correlations, and Covariances for Test No. 14
Algebra I

(Number of Items = 23; KR_{20} = .82)

	Between Schools (N = 52)			Teachers w Schools (N = 562)			Pooled w. Schools (N = 5879)		
	FINAL	STEELS	MARK	FINAL	STEELS	MARK	FINAL	STEELS	MARK
	78.80	78.64	78.70	76.21	76.03	77.26	74.21	76.39	75.04
CORRELATION \ COVARIANCE MATRIX[a,b]									
FINAL	34.69	24.71	30.00	65.51	23.35	51.18	140.60	80.52	108.17
STEELS	.57	54.82	20.84	.32	78.56	20.98	.60	127.41	68.25
MARK	.89	.49	32.74	.77	.30	62.93	.80	.53	129.76

[a]Variances are given on the diagonals, covariances above the diagonal, and correlations below the diagonal.

[b]All correlation coefficients are significant at $p < .01$.

TABLE 8

Nested Analysis of Variance of Final Test Grade, STEELS Scores, and Course Marks for Algebra I

(Kids w/ Teachers w/ Schools)

		FINAL		STEELS		MARK	
Source	d.f	M. Sqr.	F	M. Sqr.	F	M. Sqr.	F.
Schools (S)	51	2271.50	4.2	3646.69	9.5	2332.07	5.1
Tchrs. (T) w S	510	535.35	2.4	583.38	1.8	452.44	2.3
Kids w T w S	5317	219.56		208.30		197.82	

Note: All F-Ratios are significant at $p < .001$.

TABLE 9

Means, Correlations, and Covariances for Test No. 20
Pre-Calculus

(Number of Items = 27; KR_{20} = .79)

	Between Schools (N = 12)			Teachers v Schools (N = 46)			Pooled v. Schools (N = 285)		
	FINAL	STEELS	MARK	FINAL	STEELS	MARK	FINAL	STEELS	MARK
	79.90	79.96	83.27	80.73	79.80	82.98	80.40	81.14	82.80
	CORRELATION \ COVARIANCE MATRIX[a,b]								
FINAL	16.19	11.64	6.15	79.66	11.29	23.79	93.28	42.10	64.40
STEELS	.55	27.56	-3.18	.19	44.15	-4.67	.58	55.92	29.18
MARK	.56	-.22	7.50	.33	-.09	63.86	.76	.44	77.00

[a]Variances are given on the diagonals, covariances above the diagonal, and correlations below the diagonal.

[b]Underlined correlation coefficients are not significant; all others are significant at at least $p < .05$.

TABLE 10

Nested Analysis of Variance of Final Test Grade, STEELS Scores, and Course Marks for Pre-calculus

(Kids w/ Teachers w/ Schools)

		FINAL		STEELS		MARK	
Source	d.f	M. Sqr.	F	M. Sqr.	F	M. Sqr.	F.
Schools (S)	11	317.43	1.4	423.52	2.4	190.11	1.1
Tchrs. (T) w S	34	221.00	1.5	174.40	1.8	169.10	1.4
Kids w T w S	239	146.06		97.20		118.69	

Note: Underlined F-Ratios are not significant; the other F-Ratios are all significant at $p < .01$.

Marking and Grading Systems: An Example

Where do marking systems fall apart? Aside from the contribution of inappropriate (or incomplete) training, as was just discussed, there are no simple answers to this question. It is incumbent upon the educational measurement community to conduct the research necessary to identify teachers' needs and to develop the appropriate training programs. To give some idea of the extent of the problem, I have included a concrete example of a typical marking and grading system used by secondary school teachers in the Dallas school district. The example is from a specific teacher of English I, but the system is fairly common throughout the district.

As mentioned earlier, the components that go into the final calculation of the end of semester grade include, by district policy, the three, six-week cycle (report card) grades. Each of these, in turn, are determined by test scores, classwork, homework, special projects, six-week final exams, and something called "participation." The individual schools, and sometimes the teachers within the school, are allowed to determine how much weight each of these components may carry. In the example given here, the teacher (or school) determined the weights as follows:

tests: 30%
classwork: 15%
homework: 15%
participation: 10%
special projects: 10%
six-week final: 20%

The cycle data for one student in this particular teachers class are given in the exhibit on the following page. As can be seen, there were ten tests. These included several short (ten to 12 item) spelling and vocabulary reports, a couple short answer tests over some reading assignments, and two grammar tests. Since the tests as a set contributed 30 percent toward the cycle grade, each test was worth three percent. The eight classwork grades, each of which contributed 1.88% toward the cycle grade, were based primarily on work performed in a workbook and completed in class. The zero (the fifth classwork score) was given in response to the student not bringing the workbook to class. Only four homework assignments received grades; these were all exercises in the same workbook to be completed at home. The five participation scores all were were awarded in response to whether or not the student obtained a parent's signature on each of five forms (e.g. a notice of an upcoming parent/teacher association meeting). Only two grades were possible: zero or one hundred. Since there were only five of these, each score contributed two percent toward the final cycle grade. This student, for one reason or another, never obtained a parent's signature. The special project was a book report, on which this student received a grade of 92. On the six-week, cycle final exam, the student received a respectable score of 83.

Looking just at what we might consider the real "achievement" measures, i.e., the tests, the book report, and the final exam, this individual would appear to be a somewhat above average student; indeed, the unweighted average of these three components is 84. Yet the student failed this six-week cycle with a grade of 67. The reason is obvious: the

Exhibit 1

One Six-Week Cycle Grades for a Particular Student

Tests	Cls Wrk	HomeWork	Prtcptn
(30%)	(15%)	(15%)	(10%)
1. 50	1. 83	1. 85	1. 0
2. 67	2. 72	2. 61	2. 0
3. 100	3. 70	3. 52	3. 0
4. 90	4. 100	4. 48	4. 0
5. 72	5. 0	(61.50)	5. 0
6. 75	6. 64		(0.0)
7. 73	7. 45		
8. 72	8. 80		
9. 84	(64.25)		
10. 92		Project (10%)..... 92	
(77.50)		6-wk. Final (20%).. 83	

10 Tests: Worth (0.1 x 30%) = 3% each.

8 Cls Wrk scores: Worth (.125 x 15%) = 1.875% each.

4 HomeWorks: Worth (.25 x 15%) = 3.75% each.

5 Participation scores: Worth (.2 x 10%) = 2% each.

1 Project (acutally a book report): Worth 10%.

1 End of cycle final exam: Worth %20.

Six-Week Cycle Grade: 68 (FAILED)

If MIN Grade = 50, then Six-Week Cycle Grade: 74

zeros were devastating! Thus, because the student did not, or could not obtain parental signatures on notices sent home, and forgot to bring his or her workbook to class one day, he or she received a failing grade.

There are two important areas of concern, here. The first is the use of grades as a punishment. This teacher gave zeros when a student did not follow through with an assigned responsibility. Nearly all of us will agree that this is an inappropriate use of grades. More subtle, however, is the teacher's use of the "zero" grade. Most teachers, at least most of those in the Dallas school district, use zero as a valid grade for a variety of mostly punitive reasons. Typically, the zero grade is assigned in a situation where there are no achievement data. What the teachers don't see, or can't understand, however, is the fact that grades of zero carry tremendous weight. A cursory look at a few teachers grade books found no other grades in the neighborhood of zero. Generally, except for the zeros, grades were found to range from around 50 to 100. When a student does not turn in a homework assignment the student receives a zero. On the other hand, as long as the student turns in something, even if it is completely wrong, the minimum grade he or she will receive is a 50. The point I am trying to make is that while most grades for performance fall between 50 and 100 -- what I call the functional grade scale -- students sometimes receive grades of zero. These grades are way off the functional scale, and are usually awarded as a punishment. Because grades of zero are so far from the functional grade scale, they are heavily weighted; so heavily weighted, in fact, that the damage they provide is probably considerably more severe than intended. This is a concept that is not easily gotten across to teachers.

References

Airasian, P.W. Classroom assessment and educational improvement. Paper presented at the conference, Classroom Assessment: A Key to Educational Excellence, Northwest Regional Educational Laboratory, Portland, OR, 1984.

Angoff, W.H. Scales, Norms, and Equivalent Scores. Princeton, NJ: Educational Testing Service, 1984.

Carter, K. Do teachers understand the principles for writing tests? Journal of Teacher Education, 1984, 35, 57-60.

DISD, General Information Bulletin: 1986-87, Grades 7-12. Dallas: Dallas Independent School District, 1987.

Duke, J.D. Disparities in grading practice, some resulting inequities, and a proposed nev index of academic achievement. Psychological Reports, 1983, 53,1023-1080.

Ebel, R.L. Essentials of Educational Measurement (3ed ed.) Englevood Cliffs, NJ: Prentice-Hall, 1979.

Farr, R. and Griffin, M. Measurement gaps in teacher education. Journal of Research and Development in Education, 1973, 7, 19-28.

Flemming, M. and Chambers, B. Teacher-made tests: Windovs on the classroom. In W.E. Hathavay (Ed.) Testing in the Schools: Nev Directions for Testing and Measurement, No. 19. San Francisco: Jossey-Bass, 1985, 29-38.

Gullickson, A.R. Teacher perspectives of their instructional use of tests. Journal of Educatonal Research, 1984, 77, 224-246.

Gullickson, A.R. Teacher education and teacher-perceived needs in educational measurement and evaluation. Journal of Educational Measurement, 1986, 23, 347-354.

Gullickson, A.R. and Ellvein, M.C. Post hoc analysis of teacher made tests: The goodnes-of-fit between prescription and practice. Educational Measurement: Issues and Practice. 1985, 4, 15-18.

Gullickson, A.R. and Hopkins, K.D. The context of educational measurement instruction for preservice teachers: Professor perspectives. Educational Measurement: Issues and Practice, 1987, 6, 12-16.

Goldman, R.D., Schmidt, D.E., Hevitt, B.N., and Fisher, R. Grading practices in different major fields. American Education Research Journal, 1974, 11, 343-357.

Grobe, R.P, Sheehan, D.S., Sides, R.J., and Read, B.R. Final report for test development activities, 1987-88. Report No. REIS88-016. Dallas: Dallas Independent School District, 1988.

Hills, J.R. Measurement and evaluation in the Classroom (2nd Ed.). Columbus, OH: Charles E. Merrill, 1981.

Lazar-Morris, C., Polin, L, May, R., and Barry, L. A review of the literature on test use. Loss Angeles: University of California, Center for the Study of Evaluation, 1980.

Leiter, J. and Brown, J.S. Determinants of elementary school grading. Sociology of Education, 1985, 48, 166-180.

Mehrens, W.A. and Lehmann, I.J. Measurement and Evaluation in Education and Psychology (3ed ed.). New York: Holt, Rinehart, & Winston, 1984.

Meier, J.A. The Survey Tests of Essential Elements/Learner Standards (STEELS) for Secondary Schools in 1987-88. Report No. REIS88-112. Dallas: Dallas Independent School District, 1988.

Milton, O. and Edgerly, J.W. The Testing and Grading of Students. New Rochelle, NY: Change Magazine, 1976.

Milton, O., Pollio, H.R., and Eison, J.A. Making Sense of College Grades. San Francisco: Jossey-Bass, 1986.

Olson, G.H. Elementary school results on the Survey Tests of Essential Elements/Learner Standards (STEELS): 1987-88. Report No. REIS88-113. Dallas: Dallas Indepenent School District, Office of Institutional Research, 1988.

Olson, G.H. A de-facto procedure for determining cut-scores on curriculum-referenced tests. An Occassional Paper of the Office of Institutional Research. Dallas: Dallas Independent School District, 1989.

Pedulla, J.J., Airasian, P.W., and Madaus, G.F. Do teacher ratings and standardized test results of students yield the same information? American Educational Research Journal, 1980, 17, 303-307.

Prather, J.E., Smith, G., and Kodras, J.E. A longitudinal study of grades in 144 undergraduate courses. Research in Higher Education, 1976, 10, 11-24.

Salmon-Cox, L. Teachers and standardized achievement tests: What's really happening? Phi Delta Kappan, 1981, 62, 631-634.

Scates, D.E. Differences betveen measurement criteria of pure scientists and of classroom teachers. Journal of Educational Research, 1943, 37, 1-13.

Stiggens, R.J., Conklin, N.F., and Bridgeford, N.J. Classroom assessment: A key to effective education. Educational Measurement: Issues and Practices, 1986, 5, 5-17.

Thorndike, R.L. Marks and marking systems. In R.L. Ebel (Ed.) Encyclopedia of Educational Research, 4th ed. New York: Macmillan, 1969.